PRODUCT PRICE PREDICTION: Code Description

- The Code is written in PYTHON

Retail Exploration:

- The project is done using google colab, wherein the dataset had been directly read from the google drive.
- After reading the training dataset and getting the information of each dataset, it is found that 3 columns has missing values, of which the column “brand” has maximum number of missing values.
- Correlation matrix is made using seaborn and matplotlib to analyse the relationship between the features.
- Analysis is done on different features which were having impact on the target.
- In case of ‘category_name’, the column was into the main category its subsequent categories.

Retail - cleaning and training:

- Firstly the training and test dataset are joined by making a new column ‘istrain’, which depicts 1 if the observation is in training and 0 if the observation is in test dataset.
- They are joined so as to make the codes of the features same throughout the training and test dataset.
- Before joining the dataset the ‘price’ column is dropped from the training dataset.
- Then the columns with data type ‘object’ are changed to categorical data type and then turned into unique codes.
- Then the combined dataset is broken down into training and test set.
- Then the log of each variable is taken, as the output is to be calculated in RMSLE, for which lambda function is used.
- Then the model is trained using the bagging technique, Random Forest Regressor.
- The hyperparameter tuning is done using GridSearch, and the best parameters obtained are used to fine tune model using Random forest.
- A score of 0.78 is obtained using this model.
- And the model was used to predict the price of the test data and then the predicted price array was converted into series and reversing the log price.